電子カルテの病名や医薬品名を自動タグ付けする簡易システム

金子 周司[*1]　藤田 信之[*2]　大武 博[*1]

[*1]京都大学大学院薬学研究科　[*2]製品評価技術基盤機構
[*3]福井県立大学学術教養センター

A simple system for extracting medical terms from electric texts

Kaneko Shuji[*1]　Fujita Nobuyuki[*2]　Ohtake Hiroshi[*1]

[*1]Kyoto University Graduate School of Pharmaceutical Sciences
[*2]National Institute of Technology and Evaluation
[*3]Center for Arts and Sciences, Fukui Prefectural University

We have been developing an English-Japanese thesaurus of medical terms for 20 years. The thesaurus is compatible with MeSH (Medical Subject Headings developed by National Library of Medicine, USA) and contains approximately 30 thousand headings with 200 thousand synonyms (consisting of the names of anatomical concepts, biological organisms, chemical compounds, methods, disease and symptoms). In this study, we aimed to extract medical terms as many as possible from the test data by a simple longest-matching Perl script. After changing the given UTF-8 text to EUC format, the matching process required only 2 minutes including loading of a 10 MB dictionary into memory space with a desktop computer (Apple Mac Pro). From the 0.1 MB test document, 2,569 terms (including English spellings) were tagged and visualized in a color HTML format. Particularly focusing on the names of disease and symptoms, 893 terms were found with several mistakes and missings. However, this process has a limitation in assigning ambiguous abbreviations and misspelled words. The simple longest-matching strategy may be useful as a preprocessing of medical reports.

Keywords: medical terms, text tagging

1. はじめに

1993年に発足したライフサイエンス辞書（LSD）プロジェクトは、医学および生命科学領域の学生および研究者に有用なツールや辞書を提供する目的で、学術論文で用いられる専門用語を収集し、体系的な用語データベースを構築してきた。このプロジェクトでは最新の医学・生命科学用語の適切な対訳や用法を無償オンラインサービスで検索できるサービスを提供してきた[1)]。さらに最近、我々はこの対訳辞書を米国NLMが提供しているMedical Subject Headings（MeSH）互換に改良し、20万の日英同義語を3万の統制語にマッピングした医学シソーラスとして公開しており、そこには解剖学部位名、生物学名、病名・症候名、医薬品を含む化学物質名、方法および技術などが収録されている[2)]。

このようなシソーラスの現実的応用として、テキストマイニングが挙げられる。例えば日々記述される電子カルテの中から、薬物処置の結果としての症状や転帰を抽出することによって、望ましくない有害事象を早期に見いだすことが可能になるかもしれない[3)]。好ましいことに、我々が構築したシソーラスは商品名も含めた医薬品名を収録し、病名や症状名についても日本語と英語を網羅的に収録している。さらに我々は以前に、英語テキストに含まれる専門用語に対して自動タグ付けを行うPerlスクリプトを開発していた[4)]。そこで本研究では、MedNLPタスクとして与えられた模擬電子カルテを題材にして、可能な限り多くの事物の名称をテキストから抽出し、その内容を吟味することを試みた。

2. 方法

タグ付け辞書はLSDデータベースから日英同義語、統制語、カテゴリー、MeSH統制語IDを含む4列、20万行のEUCテキストとして作成した（図1）。すべての統制語は、解剖学部位（anatomy）、生物学名（biological）、病名（disease）、化合物名（molecule）、方法（method）、知識（knowledge）のいずれかに分類した。

図1 タグ付け辞書の内容

単語だけでなく、多くの複合語を収録しているLSDの利点を最大限に活かすべく、タグ付けのためのテキストマッチングは最長一致とし、速度を速めるためEUCエンコーディングで行った。また、EUCタグ辞書は同義語のバイト長で降順ソートして用いた。英単語に関しては複数形の語尾変化に対応するPerlスクリプトとした。また模擬電子カルテに含まれていた全角英数字はASCII半角英数字に変換し、Unicode独特の文字も何らかのASCII文字に可能な限り当てはめた。

タグ付けテキストを読みやすくするため、結果は標準HTML形式でclass属性をカテゴリー分類に用いて出力するようにした。これによって利用者はカテゴリー毎に画面での色合いを変更することができ、一見して用語の属性が判別できるだけでなく、Webブラウザ上でタグ付け専門用語の上にマウスをかざしたときに統制

語が表示される「マウスオーバー機能」を付与し、さらに用語をクリックした時、タグ付けした用語の意味をWebのLSD辞書サービスで調べられるように工夫した（図2）。

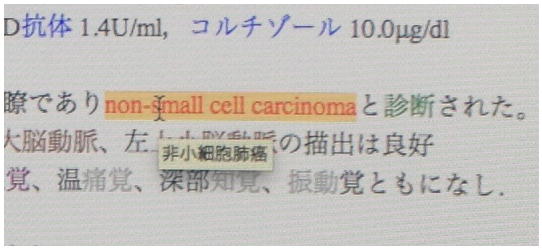

図2 マウスオーバー機能

3. 結果と考察

1121文を含む模擬電子カルテに対して、Apple Mac ProをいてUTF8からEUCへの変換を含むタグ付け処理を行ったところ要した時間は120秒であった。他のテキストで試した結果、処理速度はテキストの長さにのみ依存すると考えられた。結果として、0.1 MByteのテキストから英語スペルを含む2,569の専門用語にダグ付けがなされた。カテゴリー毎に分類すると、最も多数のタグがつけられたのは病名や症状名であり893語が抽出された。

数多くの正解タグに混じって、何パターンかの間違った、あるいは見逃された用語が散見された。最も見逃された用語はアルファベット2-3文字からなる略語である（表1）。この中にはあまり一般的でない頭字語や略称も含まれていた。この理由として、アルファベット2-3文字からなる略語は多義であることが多いため、今回のタグ付け辞書ではほとんどの略語はシソーラスから除外していたことが挙げられる。しかしながら経験的に、臨床現場において略語が頻用されることは明らかである。もし解析対象の電子カルテが特定の診療科に起因するものと判明している場合には、特に検査項目や薬物名において多用される略語を補充して使う方策は考えられるかもしれない。

表1 見逃された略語の例

Subcategory	Examples
Clinical test	T-Chol, Hb, Plt, eosino, BP, MPO, PaCO2, ALT, Cre, T-Bil, ZTT, APTT, etc.
Drug name	DIC（ダカルバジン）
	CLDM（クリンダマイシン）
	PIPC（ピペラシリン）
	PAPM/BP（パニペネム・ベタミプロン合剤）

一方、タグが適切でない代表例は、長い複合語に部分的な一致でのみタグ付けされた場合である（表2）。これらは、大きな粒度の概念がシソーラスに登録されているものの、さらに細分化された概念が未収録であることによる場合が多い。これはシソーラスの補充とともに、複合語内部での係り受け関係などを形態素解析で明らかにすることである程度まで対応できるだろう。

表2 部分一致した記述例

Partial	Compounded	More complex case
温痛覚	Murphy 徴候	眼球の黄染
顔面紅斑	心音不整	前頚部の腫脹
日光過敏	眼球結膜黄染	胆嚢軽度腫大
剥離爪	肺 MAC 症	下肺には honey comb

今回の解析で最も驚いたことは、多くの英語記述においてミススペルが検出されたこと、そして日本語においてすら誤入力によって認識されなかった用語が多かったことである。いくつかの例を表3に示すが、これらは人間の目では正しい意味として認識されているのであろう。

表3 ミススペルと誤字

In the text	Correct
predonisolone	prednisolone
theophyline	theophylline
mycobacterium abcessus	Mycobacterium abscessus
Enterococcus fecalis	Enterococcus faecalis
Klebsiella pneumonoae	Klebsiella pneumoniae
コルトコフ音	コロトコフ音
グルドパ	グルトパ（Grtpa）
クオンテェンフェロン	クオンティフェロン

4. おわりに

このタスクを通じて、与えられた英語混じりの日本語電子カルテから高速に専門用語を抽出する簡易システムが可能であることがわかった。しかしながら、多義の略語、辞書に未収録の複合概念、ミススペルや誤字などを検出するには限界があることも明らかになった。正確なテキストマイニングには形態素の構造解析などが必要と思われるが、このような簡易システムは電子カルテの表記を修正したり標準用語に修正する現実的ツールとしては有用かもしれない。

参考文献

[1] ライフサイエンス辞書プロジェクト. http://lsd.pharm.kyoto-u.ac.jp/.
[2] 金子周司,鵜川義弘,大武博,河本健,竹内浩昭,竹腰正隆,藤田信之. ライフサイエンス辞書から生命科学オントロジーへ. 情報知識学会誌, 2005;15 (2):1-10.
[3] 金子周司,天野博夫,藤田信之,大武博. 薬物有害事象AERSの医薬品名解決と薬物分類および化合物構造からの検索システム. 医薬ジャーナル 2010;46 (1):125-132.
[4] 金子周司,大武博. ライフサイエンス辞書からクリニカルインフォマティクスへ-臨床テキストからの知識発見に向けて-. 情報管理 2010;53 (9):473-479.